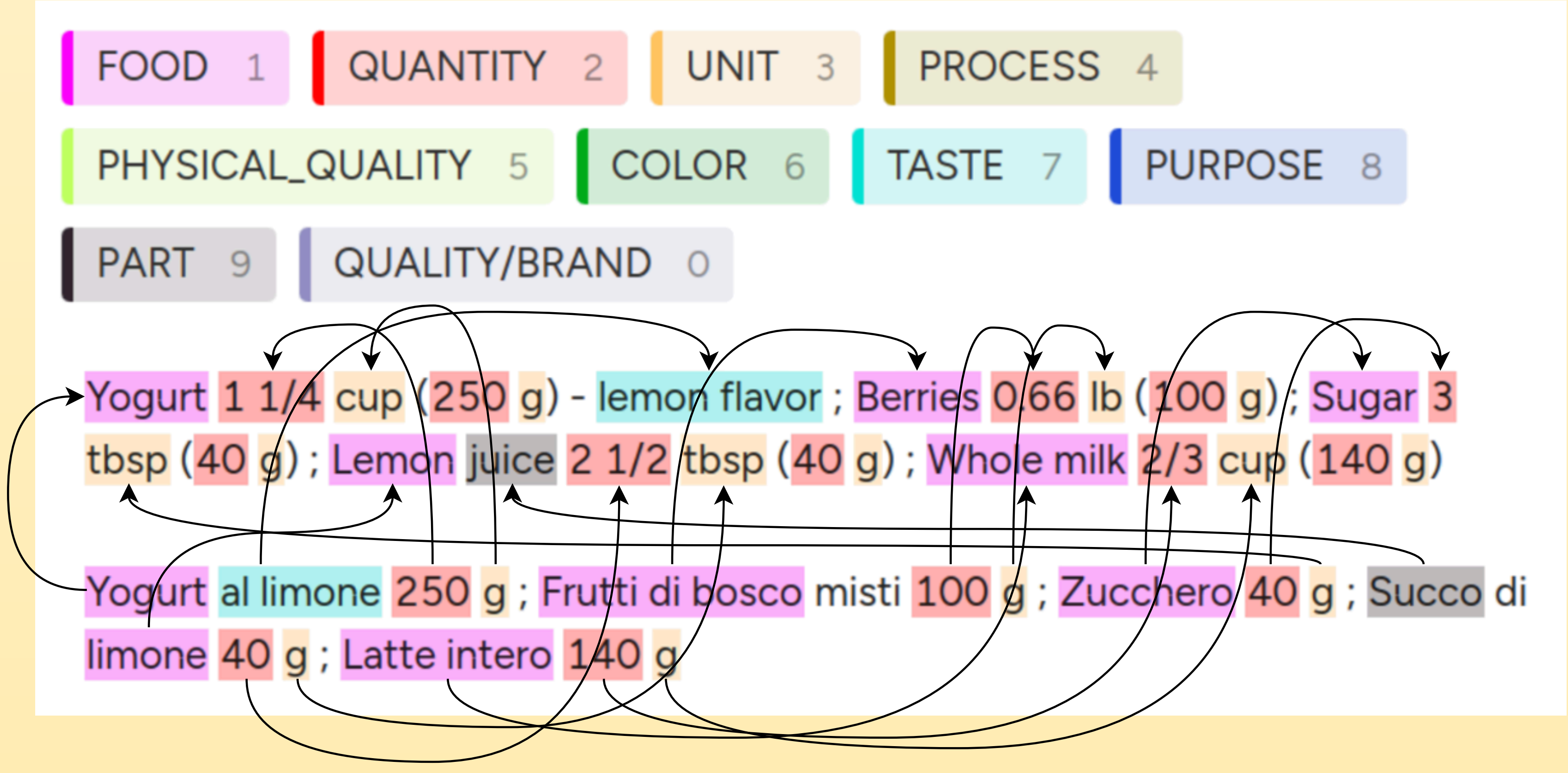
# On Cross-Language Entity Label Projection and Recognition

**Paolo Gajo and Alberto Barrón-Cedeño**
*{paolo.gajo2, a.barron}@unibo.it*
*Department of Interpreting and Translation, University of Bologna, Forlì, Italy*

| Corpus | | # Docs | Entities | | | Alignments |
|---|---|---|---|---|---|---|
| | | | en | it | es | |
| **TASTEset (EN)** | EMT | 700 | | 13,362 | | / |
| Wróblewska et al. (2022) | SMT | | 13,362 | 13,339 | 13,356 | |
| **GialloZafferano (IT > EN)**<br>Ours | | 597 | 26,631 | 20,272 | / | 9,842 |
| **My Colombian Recipes (EN > ES)**<br>Ours | | 300 | 11,551 | / | 700 | 3,565 |
| **XL-WA (EN > IT / ES)** | | 1,105 | | / | | 22,486 |
| Martelli et al. (2023) | | 1,107 | | | | 18,700 |

## Data annotation (Label Studio)

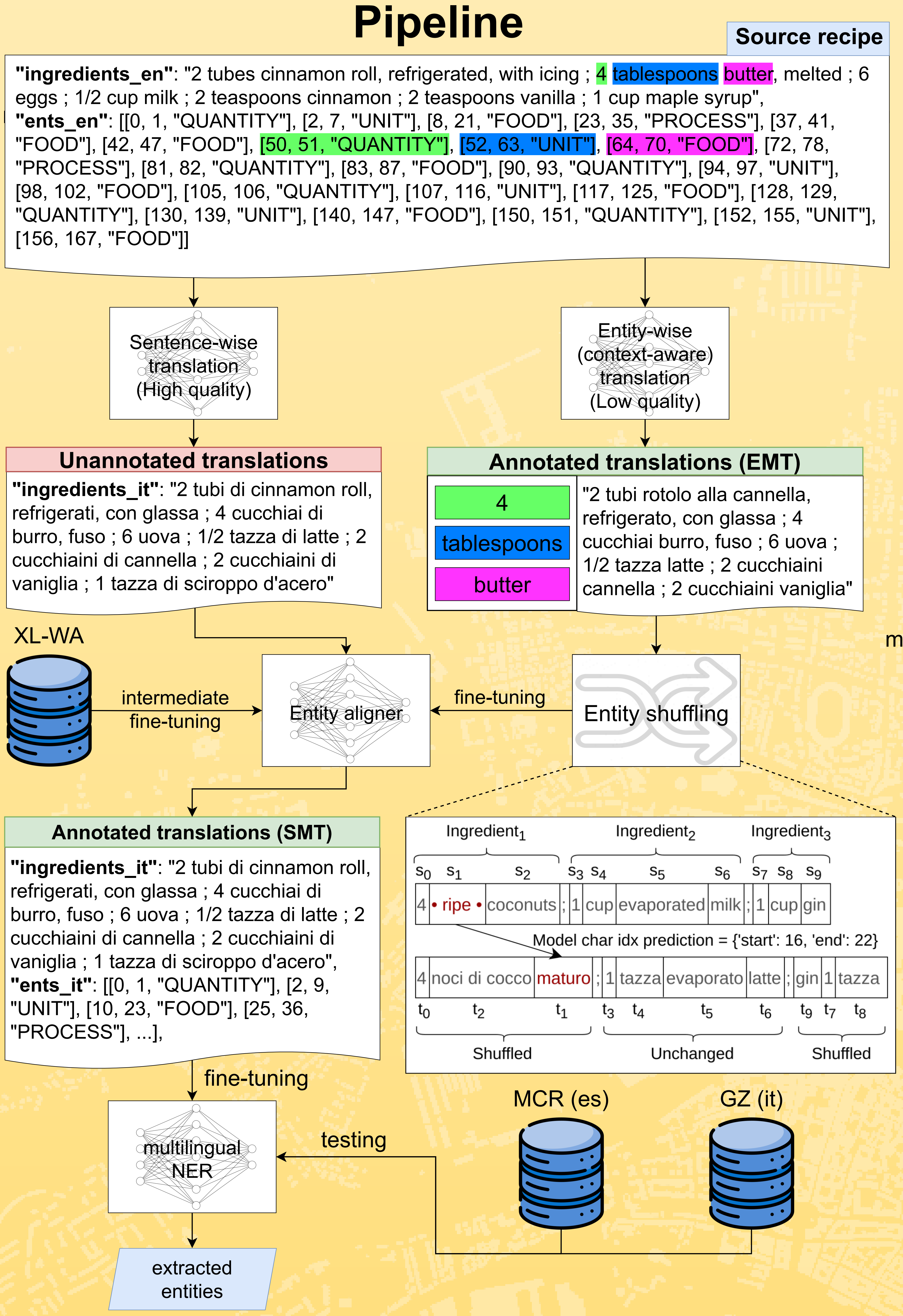


FOOD 1 | QUANTITY 2 | UNIT 3 | PROCESS 4 | PHYSICAL_QUALITY 5 | COLOR 6 | TASTE 7 | PURPOSE 8 | PART 9 | QUALITY/BRAND 0

Yogurt 1 1/4 cup (250 g) - lemon flavor ; Berries 0.66 lb (100 g) ; Sugar 3 tbsp (40 g) ; Lemon juice 2 1/2 tbsp (40 g) ; Whole milk 2/3 cup (140 g)

Yogurt al limone 250 g ; Frutti di bosco misti 100 g ; Zucchero 40 g ; Succo di limone 40 g ; Latte intero 140 g

## Pipeline



Source recipe

**"ingredients_en"**: "2 tubes cinnamon roll, refrigerated, with icing ; 4 tablespoons butter, melted ; 6 eggs ; 1/2 cup milk ; 2 teaspoons cinnamon ; 2 teaspoons vanilla ; 1 cup maple syrup",
**"ents_en"**: [[0, 1, "QUANTITY"], [2, 7, "UNIT"], [8, 21, "FOOD"], [23, 35, "PROCESS"], [37, 41, "FOOD"], [42, 47, "FOOD"], [50, 51, "QUANTITY"], [52, 63, "UNIT"], [64, 70, "FOOD"], [72, 78, "PROCESS"], [81, 82, "QUANTITY"], [83, 87, "FOOD"], [90, 93, "QUANTITY"], [94, 97, "UNIT"], [98, 102, "FOOD"], [105, 106, "QUANTITY"], [107, 116, "UNIT"], [117, 125, "FOOD"], [128, 129, "QUANTITY"], [130, 139, "UNIT"], [140, 147, "FOOD"], [150, 151, "QUANTITY"], [152, 155, "UNIT"], [156, 167, "FOOD"]]

Sentence-wise translation (High quality)

Entity-wise (context-aware) translation (Low quality)

**Unannotated translations**

**"ingredients_it"**: "2 tubi di cinnamon roll, refrigerati, con glassa ; 4 cucchiai di burro, fuso ; 6 uova ; 1/2 tazza di latte ; 2 cucchiaini di cannella ; 2 cucchiaini di vaniglia ; 1 tazza di sciroppo d'acero"

**Annotated translations (EMT)**

| | |
|---|---|
| 4<br>tablespoons<br>butter | "2 tubi rotolo alla cannella, refrigerato, con glassa ; 4 cucchiai burro, fuso ; 6 uova ; 1/2 tazza latte ; 2 cucchiaini cannella ; 2 cucchiaini vaniglia" |

XL-WA

intermediate fine-tuning

Entity aligner

fine-tuning

Entity shuffling

Ingredient$_1$
Ingredient$_2$
Ingredient$_3$
$s_0$ $s_1$ $s_2$ $s_3$ $s_4$ $s_5$ $s_6$ $s_7$ $s_8$ $s_9$
4 • ripe • coconuts ; 1 cup evaporated milk ; 1 cup gin
Model char idx prediction = {'start': 16, 'end': 22}
4 noci di cocco maturo ; 1 tazza evaporato latte ; gin 1 tazza
$t_0$ $t_2$ $t_1$ $t_3$ $t_4$ $t_5$ $t_6$ $t_9$ $t_7$ $t_8$
Shuffled
Unchanged
Shuffled

**Annotated translations (SMT)**

**"ingredients_it"**: "2 tubi di cinnamon roll, refrigerati, con glassa ; 4 cucchiai di burro, fuso ; 6 uova ; 1/2 tazza di latte ; 2 cucchiaini di cannella ; 2 cucchiaini di vaniglia ; 1 tazza di sciroppo d'acero",
**"ents_it"**: [[0, 1, "QUANTITY"], [2, 9, "UNIT"], [10, 23, "FOOD"], [25, 36, "PROCESS"], ...],

fine-tuning

multilingual NER

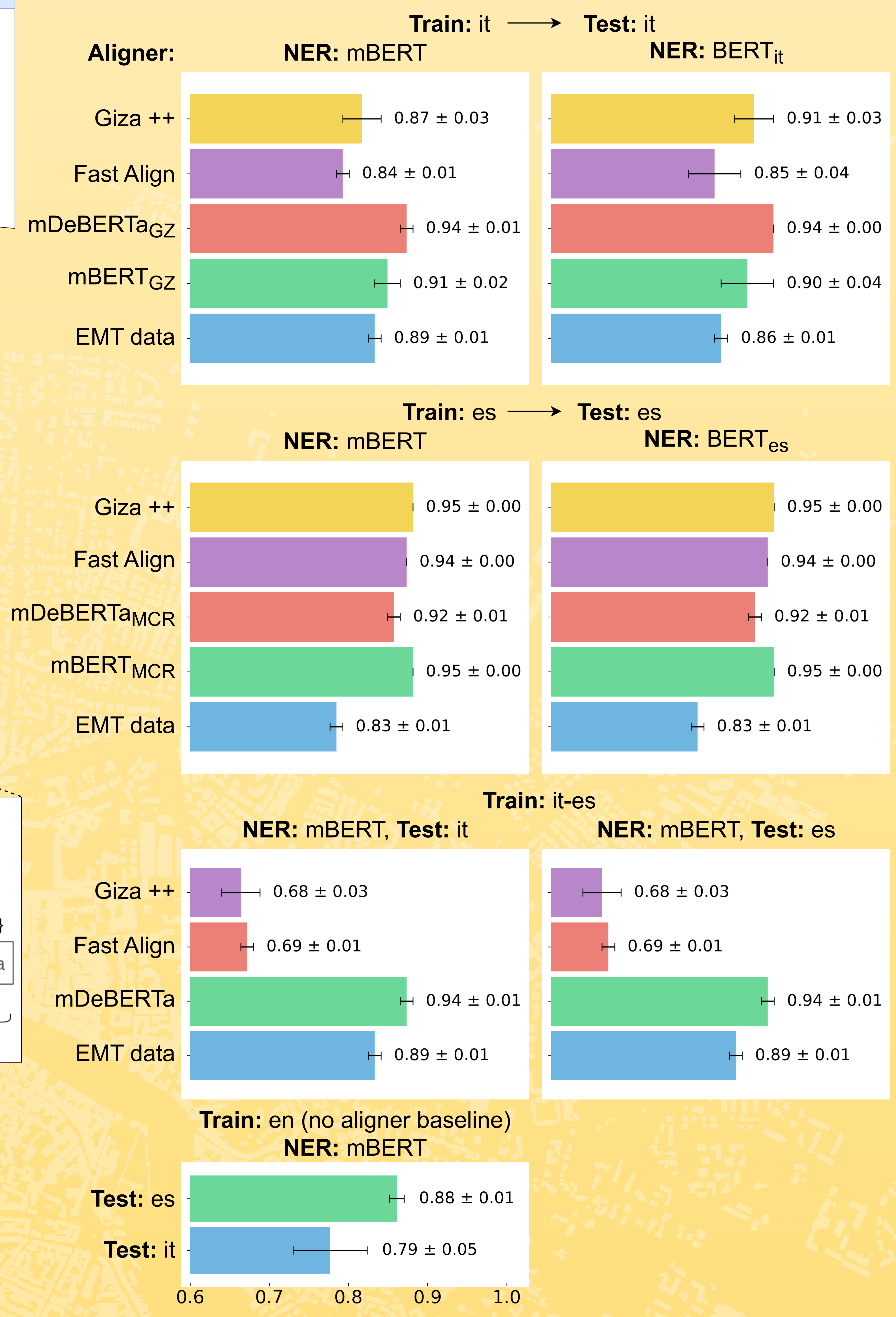
testing

MCR (es)

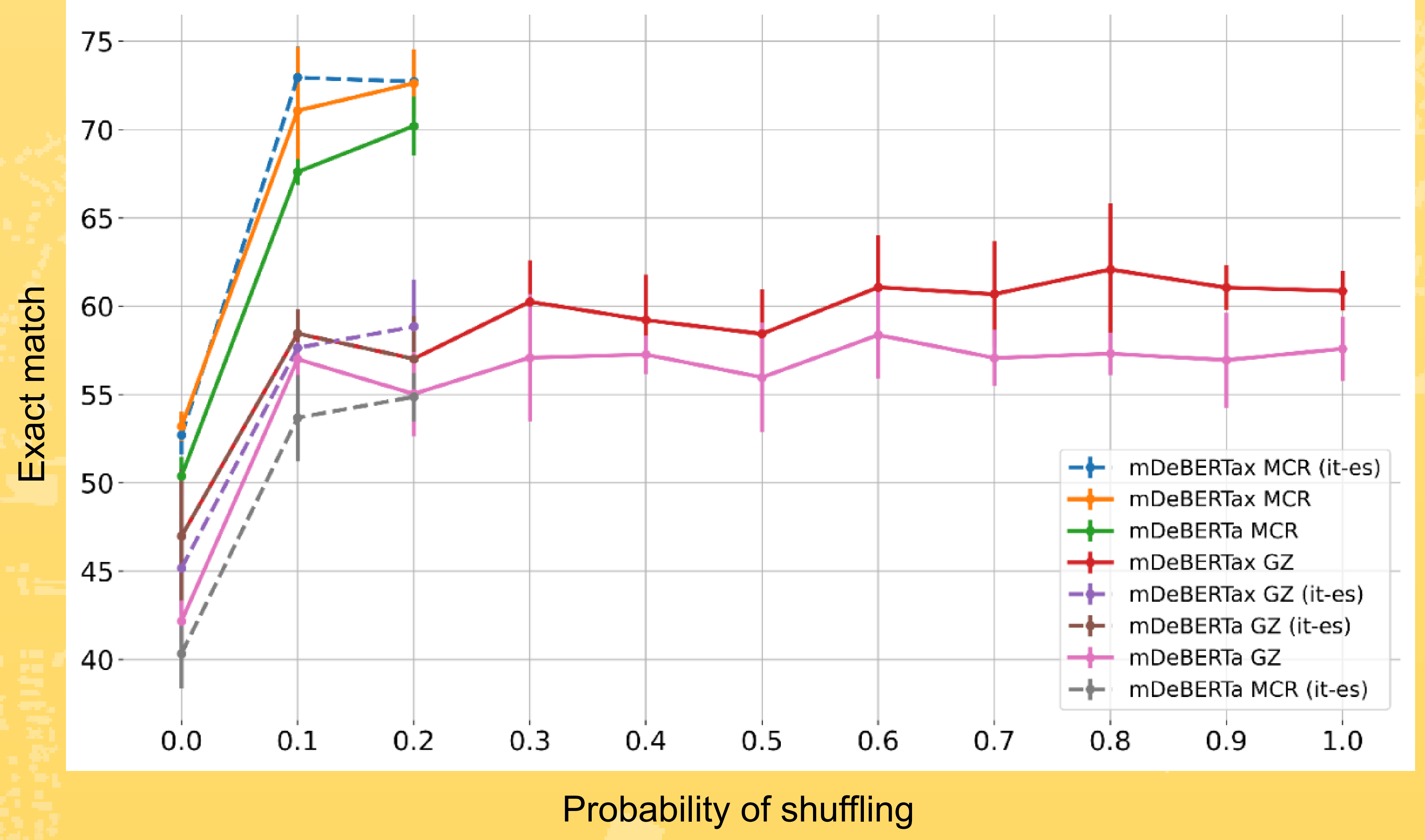
GZ (it)

extracted entities

## NER Results



## Alignment Results



## Takeaways

- Shuffling entities even with a small probability provides a large boost in alignment peformance, as the model learns not to rely on the original order of the entities.

- Generating high quality synthetic data through good entity projection models leads to better NER models, compared to simply translating entities in place. Conversely, bad alignment models lead to NER models which are worse than the entity-wise translation baseline.

- A single multilingual BERT NER model can perform as well as multiple monolingual counterparts, which means less training and inference costs.

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA A.D. 1088

CLiC it 2024

Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca